Internet of Things REU

Daway Chou-Ren

During this REU, I examined the problem of anonymous authorship analysis. We can define such a problem as follows: Given an anonymous document, can we identify which candidate’s writings samples it most closely resembles? The traditional solution is to extract stylometric features from writing samples and use statistical or machine learning algorithms to classify unknown document. Among many other applications, this has been used to classify the the Federalist Papers, Shakespeare plays, poetry, newspaper articles, and novels. But as stylometric analysis progressed, researchers moved on to tackling harder versions of the problem. Instead of analyzing print-based documents that had large samples available, well-formed writing, and few candidate authors, they began to attempt to identify shorter and noisier documents with more candidate authors. The research that followed increased the size of feature sets, incorporating features like misspellings, emoticons, document structure, and Internet lingo, allowing techniques to be applied to chat logs, forum posts, emails, tweets, and other similar document types.

Through the process of studying these problems and building an end-to-end classification system, I realized that the current research still disregards an important problem—cross-domain authorship analysis. When attempting to use authorship analysis for real-world applications such as the identification of cybercriminals or members of terrorist organizations on anonymous forums, we lack a labeled document corpus with which to compare our anonymous messages. It is far more likely that we will be able to retrieve labeled documents of another source—emails or old school assignments for example. Current techniques are unable to classify across domains and are restricted to using emails to classify emails, or private messages to classify private messages. So my project was able to further focus on studying thing cross-domain problem and assessing the feasibility of a solution.

Before approaching cross-domain analysis, I had to build an end-to-end classification system. First I read many papers in the field of authorship analysis, ranging from seminal works in the field of stylometry’s infancy to more recent and cutting edge research. I learned a great deal about existing machine learning approaches and techniques as well as about the natural language processing tools that have been used to extract stylometric features.

On a technical side, I learned how to integrate Stanford’s part of speech tagger and syntax parser into Java programs. I also learned how to manipulate csv files in R and gained a lot of experience using python for scripting tasks. I explored many different NLP toolkits for Python. In the end, my end-to-end system was composed of many R, Python, and Java programs, all wrapped in a shell script for ease of use. The system takes a corpus of labeled author data, preprocesses the data, extracts stylometric feature vectors, postprocesses and normalizes the vectors, and then runs a classifier on the labeled vectors and an unlabeled one. The classifier is trained on the labeled data and is evaluated through cross-fold validation.

The process of refining my end-to-end system was gradual. I came to realize that the more I read about other approaches, the better of a model I would be able to create. I incorporated many different previous focuses in my model and gained a deeper understanding of how syntactic, lexical, and semantic information is encapsulated in text. While my model does not perform as well as other state-of-the-art models, I have realized that the difference lies mostly in time spent in implementation. I understand the concepts behind other more advanced models very well, and given more time and more programming experience, I think I would be able to replicate such models. In any case, my model achieved results that were more than satisfactory and I can consider my model reliable for single-domain problems. For these problems, I expect the model to achieve higher than 90% accuracy for cross-validation tests. My accuracies are generally not more than the margin for error away from 100% accuracy. To get an understanding of my model’s shortcomings, I created my own unlabeled documents so that I could see what my classifier’s incorrect labels were.

Once I felt confident in my model’s performance for single-domain problems, I then tackled the problem of defining cross-domain problems. My working definition for cross-domain is that two documents may be considered to exist in separate domains when required document structure, purpose, or audience changes structural, syntactic, or lexical patterns, but not content. Using this definition, I explored classifying Facebook posts from Facebook messages.

My initial results were positive and encouraging for pursuing this problem. My model correctly classified 5/8 suspects. More work is needed to validate whether these results are significant.

Overall, the bulk of my time was spent reading and learning about current research in authorship analysis. Once I became very familiar with the state of the field, I was able to implement a working end-to-end system. I also spent a lot of time reading related materials on linguistics and sentence syntax in order to understand how part of speech tags and sentence parse patterns related to writing styles. Once I understood enough about the field, I was able to identify areas that previous work has not covered and in this way I was able to contribute a research question.